

DERWENT-ACC-NO: 1993-251901

DERWENT-WEEK: 199332

\~4~COPYRIGHT 1999 DERWENT INFORMATION LTD\~14~

TITLE: Base sequence determination for nucleic acids - includes creating graph used for analysis by plotting signals from nucleic acids against detection time

INVENTOR-NAME:

PRIORITY-DATA: 1991JP-0344354 (December 26, 1991)

PATENT-FAMILY:

PUB-NO	PUB-DATE	LANGUAGE	PAGES	MAIN-IPC
JP 05168500 A	July 2, 1993	N/A	007	C12Q 001/68

INT-CL (IPC): C12M001/00; C12Q001/68; G01N027/447; G01N033/50; G06F015/20

ABSTRACTED-PUB-NO: JP05168500A

BASIC-ABSTRACT: In base sequence determination for nucleic acid to analyse the electrophoresis pattern of nucleic acid fragments, a graph (8) is obtained by plotting signals from nucleic acid fragments against detection time. On the graph, peak areas surrounded by border lines (12) of each signal peak and base lines (13) of the graph are measured. It is determined from an index obtained from the peak areas whether each peak is formed of signals from some types of nucleic acid fragments having different base lengths (number of nucleic acid fragments in a peak); and thereby, the separation and recognition of signal peaks from continuous nucleic acid fragments.

USE/ADVANTAGE - Algorithm and hardware for base sequence determination for nucleic acid. In the conventional method of recognising peaks, base sequence determination is performed only by the recognition of maximum values, so when peaks overlap with each other it becomes impossible to separate and recognise the peaks. For this reason, for electrophoresis path length of 30 cm, one base of maximum about 500 base length can be separated and recognised, and for electrophoresis path length of 90 cm, one base of maximum about 800 base length can be separated and recognised. In this method, even where peaks cannot be separated by the conventional methods base sequence determination can be made by recognising the number of DNA fragments of different base length forming each peak. The method is useful for the sequence determination of long DNA such as human gene DNA, etc.

(19)日本国特許庁(JP)

(12) 公開特許公報(A)

(11)特許出願公開番号

特開平5-168500

(43)公開日 平成5年(1993)7月2日

(51)Int.Cl. <sup>8</sup>	識別記号	庁内整理番号	F I	技術表示箇所
C 1 2 Q 1/68	Z	8114-4B		
C 1 2 M 1/00	A	9050-4B		
G 0 1 N 27/447				
33/50	P	7055-2 J		
		7235-2 J		
			G 0 1 N 27/ 26	3 2 5 E

審査請求 未請求 請求項の数10(全 7 頁) 最終頁に続く

(21)出願番号 特願平3-344354

(22)出願日 平成3年(1991)12月26日

(71)出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72)発明者 西川 哲夫

東京都国分寺市東恋ヶ窪1丁目280番地

株式会社日立製作所中央研究所内

(72)発明者 神原 秀記

東京都国分寺市東恋ヶ窪1丁目280番地

株式会社日立製作所中央研究所内

(72)発明者 村川 克二

東京都国分寺市東恋ヶ窪1丁目280番地

株式会社日立製作所中央研究所内

(74)代理人 弁理士 小川 勝男

(54)【発明の名称】 核酸塩基配列決定方法

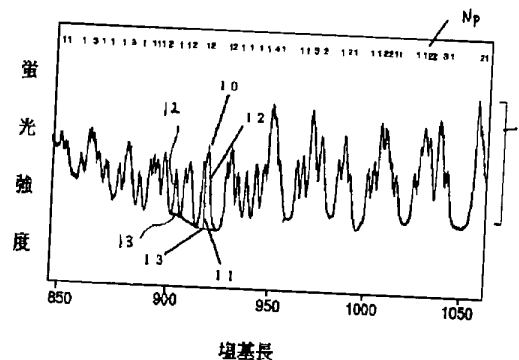
(57)【要約】

【目的】核酸断片の電気泳動パターンの解析による塩基配列決定法にて配列決定可能な最大の塩基長を大きくし、長大なDNAの塩基配列決定の効率を高める。

【構成】核酸断片からの信号を検出時刻に対してプロットしたグラフ8の上でグラフのベース線13を引き、単独のピークはその輪郭線12とベース線とで囲まれるピーク面積を、連続するピークではピーク間の谷の極小点を通る垂線11で分割されたピーク測定する。近傍のピーク面積によって該面積を規格化し、該規格化面積を各ピークを構成する1塩基ずつ長さの異なるDNA断片種の数の指標とする。

【効果】重なり合ったピークにおいても、それを構成する1塩基ずつ長さの異なるDNA断片種の数認識でき、配列決定可能な塩基配列の塩基長の限界を高めることができる。

図3.



## 【特許請求の範囲】

【請求項1】ダイデオキシ法によって生成した核酸断片の電気泳動パターンを解析する核酸塩基配列決定方法において、核酸断片からの信号を検出時刻もしくは検出時刻を単調増加で滑らかな関数によって変換した変数に対してプロットしたグラフを作成し、上記グラフ上で各々の信号ピークの輪郭線とグラフのベース線とで囲まれるピーク面積を測定し、各々のピークが1塩基長ずつ長さの異なる核酸断片種の何種類からの信号で構成されるか(ピーク中の核酸断片種の数)を前記ピーク面積から得る指標で決定して連続した核酸断片からの信号ピークの分離認識を行うことを特徴とする核酸塩基配列決定方法。

【請求項2】前記ベース線は所定の閾値より小さい極小値を有する上記グラフの谷の極小点同士を結ぶ線であり、前記閾値より大きな極小値を有する谷をはさんで連続する複数の信号ピーク群については、その谷の極小点を通る時間軸と垂直な分割線で分割し、前記ベース線、前記輪郭線及び前記分割線で囲まれる面積を個々の信号ピークのピーク面積とすることを特徴とする請求項1に記載の核酸塩基配列決定方法。

【請求項3】上記指標として、当該ピーク出現時刻の前後一定時間に検出したピーク、あるいは当該ピークの前後一定数のピークの面積情報を用いて得られる量によって規格化したピーク面積を用いることを特徴とする請求項1に記載の核酸塩基配列決定方法。

【請求項4】上記ピーク面積の規格化因子として、該ピーク出現時刻の前後一定時間に検出したピーク、あるいは該ピークの前後一定数のピークの面積のうち最小のものを用いることを特徴とする請求項3に記載の核酸塩基配列決定方法。

【請求項5】上記ピーク面積の規格化因子として、該ピーク出現時刻の前後一定時間に検出したピーク、あるいは該ピークの前後一定数のピークの面積の平均値を求め、該平均値の定数倍以上の面積を持つピークを除いて再度平均した値を用いることを特徴とする請求項3に記載の核酸塩基配列決定方法。

【請求項6】上記ピーク面積の規格化因子を求める時間範囲として、規格化するピークの出現時刻の前後5分以上の時間を用いることを特徴とする請求項3に記載の核酸塩基配列決定方法。

【請求項7】上記ピーク面積の規格化因子を求めるために用いるピークとして、規格化するピークの前後2個以上のピークを用いることを特徴とする請求項3に記載の核酸塩基配列決定方法。

【請求項8】請求項5に記載の核酸塩基配列決定方法において、ピーク面積の平均値を求めるため際に使用する定数として、1.1から2までの値を用いることを特徴とする核酸塩基配列決定方法。

【請求項9】上記各ピークに対して各ピークのピーク面積から得た指標を表示することを特徴とする請求項1に

記載の核酸塩基配列決定方法。

【請求項10】上記電気泳動パターンとしてMnバッファを使用したシーケンシング反応生成断片からのスペクトルを用いたことを特徴とする請求項1に記載の核酸塩基配列決定方法。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は核酸の塩基配列決定のためのアルゴリズム、及びソフトウェアに関する。

10 【0002】

【従来の技術】DNAの塩基配列を決定する方法として、蛍光体で標識したDNAからの蛍光を電気泳動中に実時間検出する方法が最近用いられている(バイオテクノロジー, 1988年, 6巻, pp816-821(Bio/Technology, 1988, 6, pp816-821))。ヒト遺伝子DNAなどの長大なDNAの塩基配列決定を行うためには、一回の測定で決定可能な塩基長をできるだけ大きくすることが望まれる。一回の測定で決定可能な塩基長には限界があり、この限界はポリアクリルアミドゲルを使用する電気泳動におけるDNA断片の分離限界塩基長によって決定される。

すなわち、ゲル電気泳動においては、1塩基長だけ異なるDNA断片どうしのピーク分離が塩基長が大きくなると共に困難になり、ある塩基長以上になると分離検出ができなくなる。これは、塩基長の増大に伴うピーク半値幅の減少の度合いがピーク間隔の減少の度合いに比べて小さく、ある塩基長以上になると半値幅が間隔に比べて小さくなり、隣り合ったピーク同士の分離が不可能になることによっている。1塩基の分離が可能な最大の塩基長を大きくするためには、泳動板長を長くすることと共に蛍光スペクトル中のピークの認識法の精度が重要である。従来法では、ピークの認識は基本的に極大値の認識に基づいており、泳動路長が30cmのときには約500塩基長までの1塩基分離認識、泳動路長が90cmのときには約800塩基長までの1塩基分離認識が可能である。

30 【0003】

【発明が解決しようとする課題】従来のピークの認識法では、極大値の認識のみで塩基配列の決定を行っているため、ピーク同士が重なってしまうとピークの分離認識ができなくなる。そのため、泳動路長が30cmのときには約500塩基長までの1塩基分離認識、泳動路長が90cmのときには約800塩基長までの1塩基分離認識しかできなかった。本発明の目的は、長いDNA塩基長において、ピーク同士が重なって極大値の認識ではピークの分離が不可能な場合に於ても、当該ピークが1塩基長ずつ長さの異なる核酸断片種の何種類からの信号で構成されるか(ピーク中の核酸断片種の数)を認識することによって、塩基配列決定可能な最大の塩基長を大きくし得る塩基配列決定法を提供することにある。

40 【0004】

【課題を解決するための手段】上記目的を達成するため

に、核酸断片からの信号を検出時刻もしくは検出時刻を単調増加で滑らかな関数によって変換した変数に対してに対してプロットしたグラフ上で、信号ピークの輪郭線とグラフのベース線とで囲まれるピーク面積を測定し、このピーク面積を当該ピーク中の核酸断片種の数の指標に用いることによって、連続した核酸断片からの信号ピークの実質的な分離認識を行う。

【0005】

【作用】上記手段を用いることによって以下のことが可能になる。ピーク同士が重なって極大値の認識ではピーク10の分離が不可能な場合に於ても、当該ピーク中に含まれる核酸断片種の数を認識することによって、塩基配列の決定を行うことができる。これによって、従来法にて約500塩基長までの1塩基分離が可能なピークスペクトルに対しては約700塩基長までの塩基配列決定が可能になり、従来法にて約800塩基長までの1塩基分離が可能なピークスペクトルに対しては約1000塩基長までの塩基配列決定が可能になる。このことは、塩基配列決定の決定効率を数倍高め、ヒト遺伝子DNAなどの長大なDNAの塩基配列決定に非常に有効となる。

【0006】

【実施例】以下、本発明の実施例を説明する。

【0007】(実施例1) 実施例1を図1、図2、図3、及び図4を用いて説明する。

【0008】図1はピーク面積を用いたピーク分離認識アルゴリズムのフローチャートであり、図2は93cmの泳動によって分離検出したA反応DNA断片のピークスペクトルである。図3はA反応DNA断片のピークスペクトル中である。図4はA反応DNA断片の規格化面積のグラフである。

【0009】本方法の基本的な考え方を図2を用いて説明する。図2のグラフはA反応DNA断片のピークスペクトル7を塩基長を横軸にして描いたものである。DNA断片はM13mp8ファージを鋳型にしてシーケンシングA反応を行ったものである。DNA断片の末端をフルオレセインイソチオシアネートによって標識し、アルゴンレーザーの励起で蛍光スペクトルを得た。泳動距離は30cmであり、蛍光スペクトルからは500塩基長までは隣り合ったピーク同士が分離して検出されており、正確な塩基配列決定が可能である。500塩基長以上になるとピークの幅が大きくなり隣り合ったピーク同士が分離しなくなる。従って、500塩基長以下のピークは全て単一の長さのDNA断片からのピークであるが、500塩基長以上のピークには、塩基長が1塩基ずつ異なる2個以上のピークが重なり合って単一のピークとして観察されるものが存在する。図2からわかるように、500塩基長以下の単一のピークはピーク強度が塩基長とともに非常に一様に変化しており、平均曲線からの変動は約15%以内である。この一様性は、シーケンシング反応においてMnを用いた反応バッファーを使用したことによって実現

されている(文献; ジャーナル オブ バイオロジカル ケミストリー, 1990年, 265巻, pp 8322-8328. (J. Biol. Chem., 1990, 265, pp8322-8328.))。ピーク強度が一様に変化するという事は、ピーク面積が一様に变化することを意味する。故に、n個重なり合ったピークの面積はその近傍の単一ピーク面積の約n倍になる。従って、逆にピーク面積を測定しモニターしていけば、分離されていない各ピークが実際には何種類の塩基長の断片種の信号から構成されるかを推定することができる。

10 【0010】次に、ピーク分離認識アルゴリズムの詳細を図1のフローチャートを用いて説明する。本方法は、各ピーク群の分割1、分割面積の測定2、ピーク面積の規格化因子の計算3、各ピーク面積の規格化4、ピーク個数の推定5、ピーク個数の表示6から成る。

【0011】まず、各ピーク群の分割1を行う。各ピーク群の分割は図3に示したように行う。図3は93cmの泳動によって分離検出したA反応断片のピークスペクトル8であり、850塩基長から1063塩基長までのピークを含んでいる。横軸は塩基長で表示しているが、実際の検出信号は泳動開始からの検出時刻を変数としてプロットされる。さらに、その検出時刻をある単調増加で滑らかな関数によって変換した変数に対して信号をプロットすることにより、図示したように横軸をリニアな塩基長軸とみなすことができる。各ピークの上には各ピークに対応する1塩基長ずつ長さの異なる塩基種の数 $N_p$ を表示している。図3からわかるように、塩基長が一つ異なる2種の断片からの信号は分離されてない。図には示していないが、同条件の泳動で800塩基長までのピークスペクトルでは、連続したピークが分離して検出される。一方、800塩基長を越えるピークスペクトルについては、まずピークスペクトルの谷のうち極小値が所定の閾値より小さい谷の極小点をつないでベース線13とする。これらの極小点の間に極小値が上記閾値を越える小さな谷がある場合には、つまりピーク群が隣接して連なっているときには単独のピーク毎への分割が必要となる。そこで、ピーク群中の谷の底(極小点)を通る垂線11を引きピーク群を分割する。なお、上記の閾値はプロットされたグラフに応じて、所定の区域がごとくに定めるのが好ましい。

40 【0012】分割面積の測定2は、この垂線11、ピークの輪郭線12、及びベース線13で囲まれる面積の測定によって行う。もともと単独のピークについては、ピークの輪郭線12とベース線13とで囲まれる面積が測定される。以上の操作はC、G、Tの反応断片からそれぞれ得られたピークについても行う。

【0013】次に、ピーク面積の規格化因子の計算3を次のように行う。まず、着目するピークを含めて前後一定数(例えば全部で5個)のピークの面積を平均し、この平均値をPとする。この平均の中から $N_p=1$ のピーク、つまり1種類の塩基長の断片の信号ピークを抽出す

る、ピーク面積がPの定数倍(例えば1.1倍)以上のピークを除外して、再度平均する。これによって、ピーク面積の規格化因子P'が求まる。この平均操作においては、平均の範囲をピーク個数ではなく、着目するピークの前後一定時間(例えば5分間)ととってもよい。また、規格化因子としては、前後一定数あるいは一定時間中のピークの面積が最小のものを採用してもよい。

【0014】各ピーク面積の規格化4は、各々のピークの規格化因子P'で各ピークの面積を除することによって行われる。DNA断片種の数の推定5は、規格化面積を四捨五入することによって行われる。図4に図3のピークスペクトルから上述の方法によって規格化ピーク面積を求め、塩基長に対してプロットしたグラフを示す。N<sub>p</sub>の値によって異なったシンボル(14; N<sub>p</sub>=1、15; N<sub>p</sub>=2、16; N<sub>p</sub>=3、17; N<sub>p</sub>=4)で表示した。N<sub>p</sub>=1のピークの規格化ピーク面積をS<sub>1</sub>、N<sub>p</sub>=2のピークの規格化ピーク面積をS<sub>2</sub>、N<sub>p</sub>=3のピークの規格化ピーク面積をS<sub>3</sub>、N<sub>p</sub>=4のピークの規格化ピーク面積をS<sub>4</sub>と表せば、S<sub>1</sub>、S<sub>2</sub>、S<sub>3</sub>、S<sub>4</sub>の平均値はそれぞれ1.0、2.0、3.0、3.95である。S<sub>1</sub>、S<sub>2</sub>、S<sub>3</sub>、S<sub>4</sub>のそれぞれ1、2、3、4からのずれの幅はそれぞれ+0.15~-0.15、+0.3~-0.2、+0.35~-0.25、-0.1であるから、S<sub>1</sub>、S<sub>2</sub>、S<sub>3</sub>、S<sub>4</sub>を四捨五入した値は全ピークについてそれぞれ、1、2、3、4になり、間違いなくピーク数の認識が可能である。C、G、Tの反応断片から得られたピークについてもそれぞれ同様なことを行えば、約1050塩基長までの塩基配列が正確に決定されることになる。

【0015】以上の方法によって求めたDNA断片種の数の表示6は、図2中のピークスペクトル中に表示したように行われる。泳動距離が30cmの場合にも、規格化面積を用いたピーク分離認識法を適用すれば、約700塩基長までの塩基配列決定が正確に行えるようになる。

【0016】

【発明の効果】本発明によれば、ピーク同士が重なって極大値の認識ではピークの分離が不可能な場合に於て

も、各ピークを構成する1塩基長ずつ長さの異なるDNA断片種の数を確認することによって、塩基配列の決定を行うことができる。これによって、約500塩基長までの1塩基分離が可能なピークスペクトルに対しては約700塩基長までの塩基配列決定が可能になり、約800塩基長までの1塩基分離が可能なピークスペクトルに対しては約1000塩基長までの塩基配列決定が可能になる。このことは、塩基配列決定の決定効率を数倍高め、ヒト遺伝子DNAなどの長大なDNAの塩基配列決定に非常に有効となる。

【図面の簡単な説明】

【図1】本発明の実施例1の説明図で、ピーク面積を用いたピーク分離認識アルゴリズムのフローチャートである。

【図2】A反応DNA断片のピークスペクトルである。

【図3】93cmの泳動によって分離検出したA反応DNA断片のピークスペクトルである。

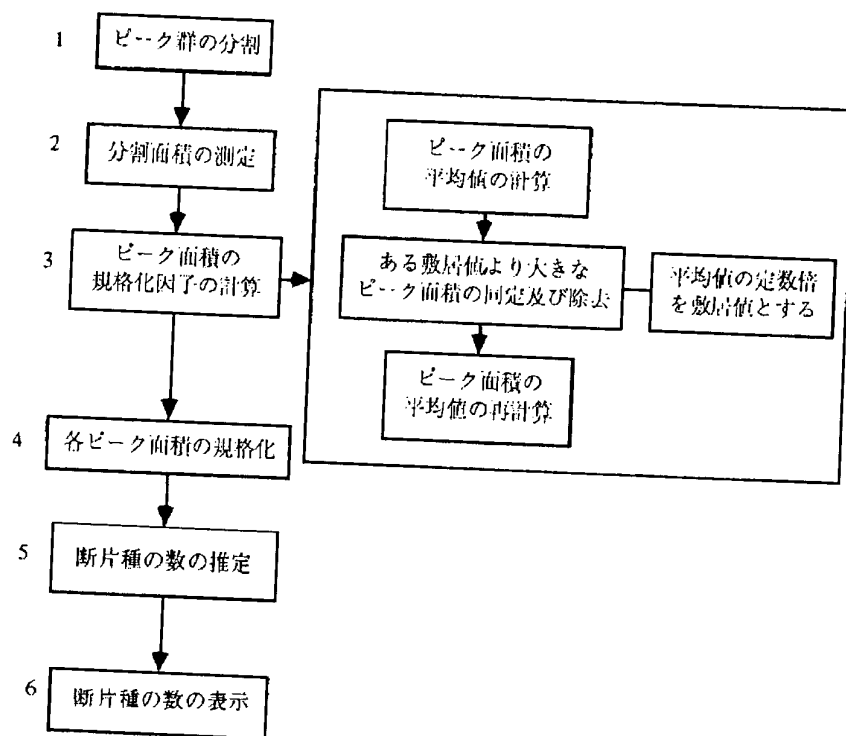
【図4】A反応DNA断片の規格化面積のグラフである。

【符号の説明】

- 1...各ピーク群の分割、
- 2...分割面積の測定、
- 3...ピーク面積の規格化因子の計算、
- 4...各ピーク面積の規格化、
- 5...DNA断片種の数の推定、
- 6...DNA断片種の数の表示、
- 7、8...A反応DNA断片のピークスペクトル、
- 10...ピーク群、
- 11...ピーク群中の谷の底を通る垂線、
- 12...ピークの輪郭線、
- 13...ベース線、
- 14...N<sub>p</sub>=1のピーク、
- 15...N<sub>p</sub>=2のピーク、
- 16...N<sub>p</sub>=3のピーク、
- 17...N<sub>p</sub>=4のピーク。

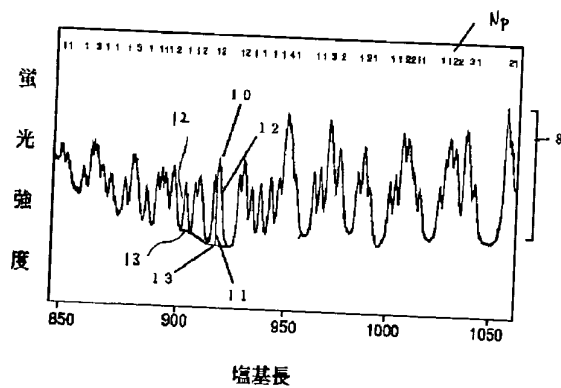
【図1】

図1



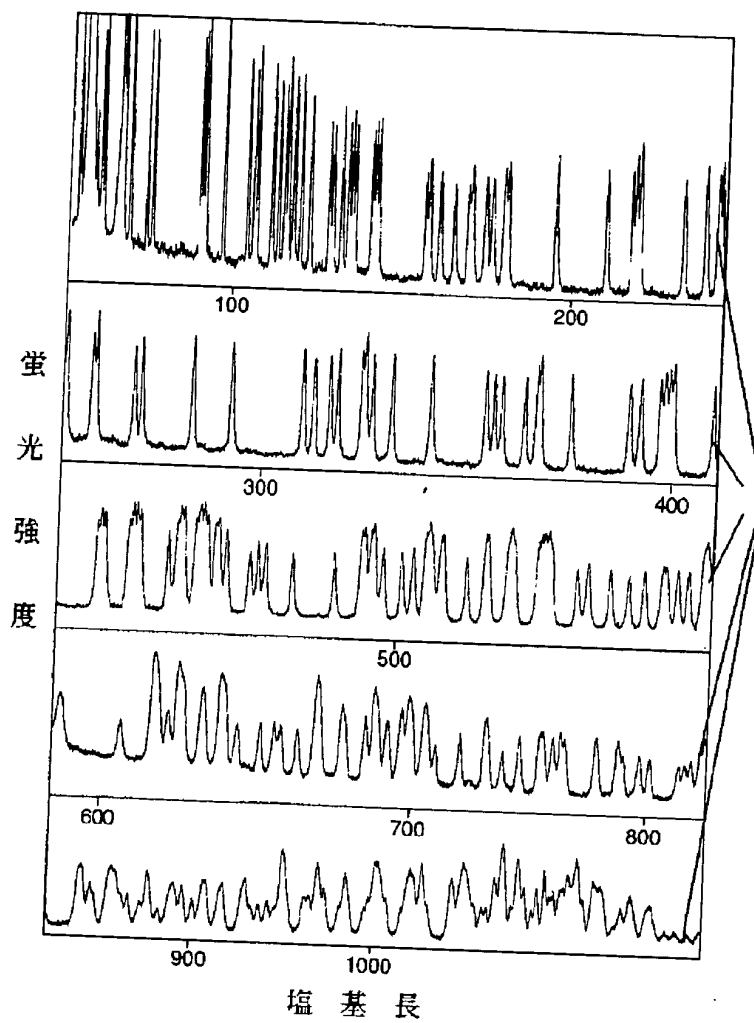
【図3】

図3.



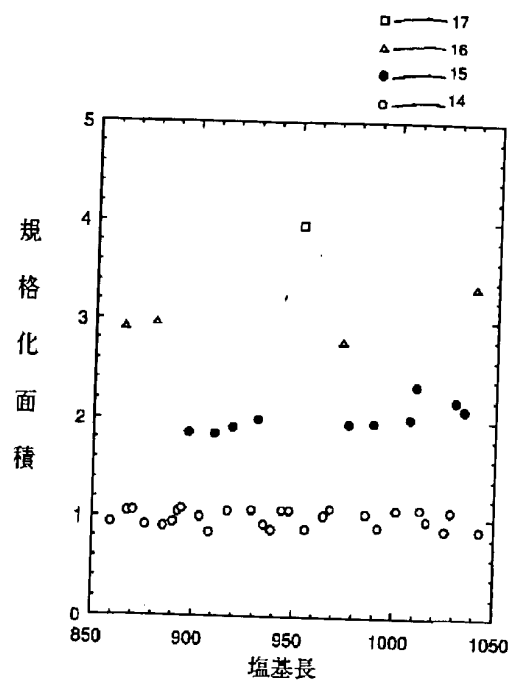
【図2】

図2



【図4】

図 4



フロントページの続き

(51)Int. Cl.<sup>5</sup>

G 0 6 F 15/20

識別記号

庁内整理番号

F I

F 7218-5L

技術表示箇所

L Number	Hits	Search Text	DB	Time stamp
1	86	sequencing.clm. and peak.clm.	USPAT; US-PGPUB	2002/03/22 09:19
4	1745	sequenc\$4 and peak and (scale or time or migration) and (reference or known) and polynomial	USPAT; US-PGPUB	2002/03/22 09:42
7	301	(sequenc\$4 and peak and (scale or time or migration) and (reference or known) and polynomial) and trace	USPAT; US-PGPUB	2002/03/22 09:37
10	23	((sequenc\$4 and peak and (scale or time or migration) and (reference or known) and polynomial) and trace) and (DNA or polynucleotide or nucleic adj acid)	USPAT; US-PGPUB	2002/03/22 09:38
13	22	((((sequenc\$4 and peak and (scale or time or migration) and (reference or known) and polynomial) and trace) and (DNA or polynucleotide or nucleic adj acid)) not (sequencing.clm. and peak.clm. )	USPAT; US-PGPUB	2002/03/22 09:38
16	2	sequenc\$4 and peak and (scale or time or migration) and (reference or known) and polynomial	EPO; JPO; DERWENT; IBM_TDB	2002/03/22 09:42
21	1106	sequenc\$4 and peak and (scale or time or migration)	EPO; JPO; DERWENT; IBM_TDB	2002/03/22 09:43
26	2	sequenc\$4 and peak and (scale or time or migration) and polynomial	EPO; JPO; DERWENT; IBM_TDB	2002/03/22 09:43
31	50	sequenc\$4 and peak and (scale or time or migration) and (DNA or polynucleotide or nucleic adj acid)	EPO; JPO; DERWENT; IBM_TDB	2002/03/22 09:43
36	29	(sequenc\$4 and peak and (scale or time or migration) and (DNA or polynucleotide or nucleic adj acid)) not us.pc.	EPO; JPO; DERWENT; IBM_TDB	2002/03/22 09:44